AFRL-SR-AR-TR-04-

0537

REPORT DOCUMENTATION PAGE

1. AGENCY USE ONLY	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	8/31/04	Final Report: 12/01/00 – 05/31/04

4. TITLE AND SUBTITLE Rapidly Building High-Performance Information Agents	5. FUNDING NUMBERS F49620-01-1-0053
6. AUTHORS Craig A. Knoblock	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) University of Southern California / Information Sciences Institute 4676 Admiralty Way, Suite 1001 Marina del Rey, CA 90292	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Air Force Office of Scientific Research 4015 Wilson Blvd Arlington, VA 22203-1954	10. SPONSORING/MONITORING AGENCY REPORT NUMBER
11. SUPPLEMENTARY NOTES UM	
12A. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.	12B. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This research project addressed the problem of rapidly constructing high-performance information agents. In this work, we made three significant advances. First, we developed a technique called Co-testing to address the problem of generating high accuracy wrappers to extract data from online sources. Second, we developed a method called speculation execution for improving the performance of streaming dataflow execution systems used in information agents. Third, we developed a wizard-based approach to interactively and rapidly constructing information agents.

14. SUBJECT TERMS Information agents, machine learning, plan execution, co-testing, speculative execution	15. NUMBER OF PAGES 17 pages
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UNLIMITED

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. Z39-18
289-102

20041028 085

2004 Final Performance Report

Rapidly Building High-Performance Information Agents

USAF, Air Force Office of Scientific Research
Award Number: F49620-01-1-0053
Period of Performance: 12/01/00 – 5/31/04

Craig A. Knoblock (PI)
University of Southern California
Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292-6695
Phone: 310-448-8786
Fax: 310-822-0751
Knoblock @isi.edu

Status of the Effort:

This research project addressed the problem of rapidly constructing high-performance information agents. In this work, we made three significant advances. First, we developed a technique called Co-testing to address the problem of generating high accuracy wrappers to extract data from online sources. Second, we developed a method called speculation execution for improving the performance of streaming dataflow execution systems used in information agents. Third, we developed a wizard-based approach to interactively and rapidly constructing information agents.

Accomplishments/New Findings:

In this section we describe our contribution on wrapper learning, agent plan optimization, and rapidly constructing information agents. This project resulted in two Ph.D. theses on these topics, which are available from http://www.isi.edu/~knoblock:

- Ion Muslea.
  Active Learning with Multiple Views.
  Department of Computer Science, University of Southern California, 2002.

- Greg Barish.
  Speculative Plan Execution for Information Agents.
  Department of Computer Science, University of Southern California, 2003.

Learning for Wrapper Generation

Labeling training data for learning algorithms is a tedious, error prone, and time-consuming process. Active learning addresses this issue by detecting and asking the user to label only the most informative examples in a domain. In this project, we developed a technique called Co-Testing [Muslea *et al.*, 2000], an active learning technique for domains with multiple *views*; i.e., domains with disjoint sub-sets of features, each of which is sufficient for learning. Co-Testing is a two-step iterative algorithm that (1) uses the few available labeled examples to learn a hypothesis in each view and (2) queries (i.e., asks the user to label) examples on which the views predict a different label. Such queries are highly informative because they correct mistakes made by one of the views: whenever the views disagree, at least one of them must be wrong.

Co-Testing was successfully applied to wrapper induction [Muslea *et al.*, 2000], an industrially important application. In wrapper induction the goal is to learn rules that extract the relevant data from collections of Web pages that share the same underlying structure; e.g., extract the book titles and prices from `amazon.com`. For wrapper induction, Co-Testing uses two views: the sequences of tokens that precede and follow the extraction point, respectively. The extraction rules learned in these views are finite automata that consume an item's prefix or suffix within the page, respectively.

Co-testing with Strong and Weak Views

The main limitation of existing Co-Testing algorithms [Muslea *et al.*, 2000; 2002a] is that they are designed to use only views that are adequate for learning, thus being unable to also exploit imperfect views that would permit a faster convergence to the target concept. To address this problem, we extended the multi-view learning framework by introducing the idea of learning from *strong* and *weak* views. By definition, a *strong view* consists of features that are adequate for learning the target concept; in contrast, in a *weak view* one can only learn a concept that is more general or specific than the target concept. We introduced a novel algorithm, *Aggressive Co-Testing*, that exploits both strong and weak views without additional data engineering costs. We also described a case study on wrapper induction, which shows that Aggressive Co-Testing clearly outperforms state-of-the-art algorithms. We used a collection of 33 difficult extraction tasks to show that using the weak view dramatically reduces the need for labeled data: compared with existing state of the art active learners, our novel algorithm requires between 45% and 81% fewer labeled examples.

Adaptive View Validation

In practice, Co-Testing clearly outperforms the other wrapper induction approaches on the vast majority of the extraction tasks. However, there are scenarios in which Co-Testing is not the most appropriate algorithm to be used. As mentioned in the previous report, we conducted a preliminary empirical study that showed Co-Testing is highly successful only when both types of extraction rules are equally well suited for the extraction task. However, in practice, one does not know before hand whether or not the two types of extraction rules are appropriate for a new, unseen extraction task.

In order to cope with this problem, we formalized the concept of view validation and introduced an adaptive view validation algorithm. This novel algorithm generalizes and improves our previous work on manually finding heuristics that are useful at predicting whether or not one should apply Co-Testing to a new, unseen task.

The problem of view validation can be described as having to predict whether or not the multi-rule approach is appropriate for a new, unseen task. To address this issue, we introduce an Adaptive View Validation algorithm that learns to predict whether Co-Testing is the most appropriate algorithm for a new task. More precisely, the Adaptive View Validation algorithm uses the experiences acquired while solving past extraction tasks to predict the most appropriate algorithm for a given task.

Our Adaptive View Validation algorithm takes as input a set of extraction tasks that are labeled by the user as being appropriate or inappropriate for Co-Testing. Then the view validation algorithm uses features such as the complexity of the learned rules or their error rates on the training data to learn a classifier that predicts whether or not Co-Testing is appropriate for a new task. Our empirical results for both wrapper induction and the (related) problem of text classification show that the Adaptive View Validation makes highly accurate predictions based on a modest amount of training data. These results clearly outperform the hand-written heuristics described in our previous report: Adaptive

View Validation reaches an accuracy of 92%, while the older rules were only 76% accurate.

Optimizing the Execution of Information Agent Plans

The performance of Web information gathering plans can suffer because of I/O latencies associated with the remote sources queried by these plans. A single slow Web source can create a bottleneck in an entire plan and lead to poor execution time. When a plan requires multiple queries (either to the same source or to multiple sources), performance can be even worse, where the overhead is a function of the slowest sequence of sources queried.

When multiple queries are required, speculative plan execution (Barish and Knoblock 2002) can be used to dramatically reduce the impact of aggregate source latencies. The idea involves using data seen early in plan execution as a basis for issuing predictions about data likely to be needed during later parts of execution. This allows data dependency chains within the plan to be broken and parallelized, leading to significant speedups.

The Theseus plan executor is implemented as a virtual dataflow machine. Execution involves the streaming of information into the plan, the decentralized and parallel processing of that information and its intermediate results by these operators, and finally the output of one or more result streams. Each operator in the plan is represented at execution time by one or more threads. Communication between producer and consumer operator threads is asynchronous; a producing operator can deposit information into the queue of a consuming operator and thus not be forced to wait for the consumer to finish its current work before continuing with its own execution. This asynchronous streaming, the dataflow nature of execution, additional support for concurrent transactions, and the distributed retrieval of data from multiple sources enables the Theseus executor to realize four distinct types of parallelism at run-time.

Though these four types of parallelism enable plan execution to be fast, producing information as soon as possible, execution time can still be limited by the inefficiency of one site. Consider a simple plan that retrieves a list of popular restaurants from one site and then retrieves details about those restaurants from another site. If the former site is slow, overall plan execution will be slow, no matter how fast the latter site turns out to be. At the same time, it is often the case – especially in plans that monitor information sources – that data dependencies can be cached or predicted. For example, it may be possible to predict the list of popular restaurants based on prior executions. Given the potential to predict intermediate plan data during execution and the I/O-bound nature of execution, it can be highly profitable to engage in speculative execution.

We have made two contributions related to speculative execution. First, we developed techniques to automatically augment a plan for speculative execution (Barish and Knoblock, 2002a). We developed an algorithm, called SPEC-REWRITE, that rewrites a

dataflow-style information agent plan into one capable of speculative execution. The heart of the algorithm relies on identifying the most expensive path (MEP) of a plan and rewriting the plan so that the consumers of costly operators along that path are speculatively executed. Once rewritten, the MEP of the plan is again identified and further refinement for speculative execution is attempted – a process that continues until no further refinement is possible. One of the benefits of the SPEC-REWRITE algorithm is that it supports *cascading speculation* – the speculation of future operators based on the speculation of prior operators – possible. This allows plan execution speedups to be maximized: speedups equal to the length of the longest data dependent data flow in a plan.

A second contribution to our speculative execution approach is a technique for learning how to speculate about data (Barish and Knoblock, 2002b). To maximize the impact of speculative execution on plan performance, a good value prediction strategy is required. The basic problem involves being able to use some hint h as the basis for issuing a predicted value v. Caching is one possible solution; we can note that particular hint h_x corresponds to a particular value v_y so that future receipt of h_x can lead to a prediction of v_y. However, caching has two disadvantages: it is not space-efficient since it requires the storage of all prior hint/value mappings), and it is not always applicable during execution since it only allows predictions to be made upon previously seen hints. In contrast, two other types of predictors, formed using standard machine learning techniques, can often be used to address the data speculation problem with better space efficiency and wider applicability at runtime.

The first type of predictor is formed through decision tree learning, which allows hints to be *classified* into predicted values. Decision trees are effective because they enable us to issue intelligent predictions about recurring as well as new hints, accomplishing the latter by learning which attributes of the hint are the most informative and ranking them accordingly. For example, decision trees can be used to predict that if (4676 Admiralty Way, Marina del Rey, CA) is in zip code 90292 then (4680 Admiralty Way, Marina del Rey, CA) is also in the same zip code. This prediction can be made without having previously seen that particular address.

A second type of predictor is formed through transducer learning, which allows hints to be *translated* into predictions through use of a finite state device known as a subsequential transducer. Transducers are advantageous because they are very space efficient (a finite state machine is learned, not a list of hints and values) and because they can issue predictions given new hints, and also issue new predictions (i.e., predictions not previously issued). For example, if an agent plan contains a source that simply acts as an encoding function (such sources are commonly encountered when integrating Internet data), such as one that translates "Los Angeles" to:

"http://www.weather.com/lookup.cgi?city=los_angeles",

a transducer can be used to learn that function. To combine the approaches of decision trees, transducers, and caching, we developed an algorithm called RETROSPECT that integrates the incremental learning of each of these types of predictors. RETROSPECT

learns the type of predictor that best suits the relationship between past hints and actual values.

Rapidly Constructing Information Agents

We developed a question-answering approach where a user without any programming skills can build information agents by simply answering a series of questions. These resulting agents can perform fairly complex tasks that involve retrieving, filtering, integrating and monitoring data from online sources. The goal of the question-answering approach is to create the workflow in Figure 1, which can later be converted into an executable agent, by asking a user some simple questions. The challenge is to decide what questions to ask and the order of the questions that we should ask the user. Our approach is to impose a hierarchical structure on the web sources in a form of a tree as shown in Figure 2. The lowest level is the agent level where we have agents that can extract the data from the web site. Domain level and service level are abstract levels that we introduce so we can group agents based on domains and services. The output level can be mapped to the output node in Figure1.

Based on this structure, we can derive the set of questions to ask on each node based on the level of the tree. To determine the order of the questions, we use the post order traversal of the node in the tree. As the user answers each question, the workflow in the data gathering part in Figure 1 will be generated incrementally. By the time we reach the output node of the post order traversal, the workflow in the data gathering part will be completed. For the data monitoring part, the user will be asked to select one of the seven available monitoring conditions. Based on the user choice, the workflow in the data monitoring part will be generated automatically, so the user will be sheltered from details of the query and the configuration of the database.

We evaluate how well the Agent Wizard works by using it to build a set of agents in the flight travel domain based on the Travel Elves. The Travel Elves is an application suite that let users search and monitor for information about flights that most air travelers find useful. The Travel Elves contains nine agents and it took four programmers roughly four days to implement the whole suite. To evaluate the Agent Wizard, we had two users build nine agents using the Agent Wizard that are functionally equivalent to the nine agents in Travel Elves. The first user is an expert user who knows the Agent Wizard well. The second user is one of the four programmers who implemented the Travel Elves. Using the Agent Wizard, the entire set of agents can be implemented by both users in under 35 minutes.

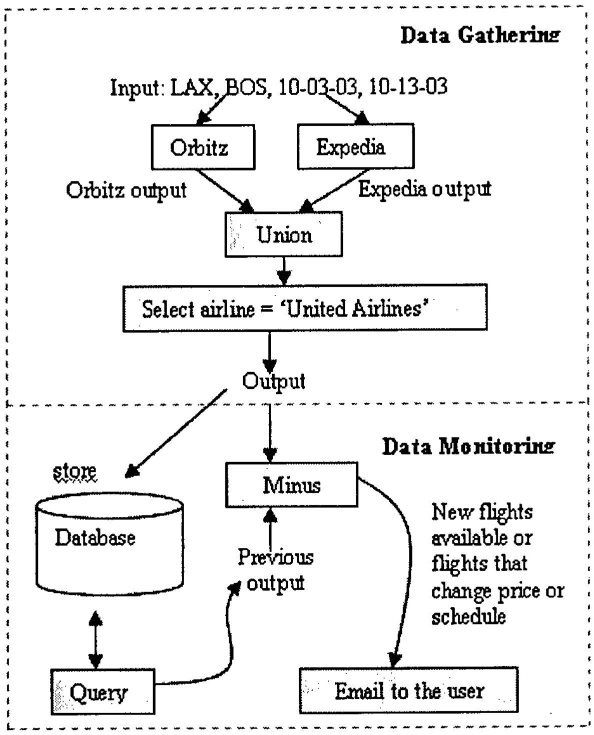

Figure 1: The workflow for our PriceMonitor Agent

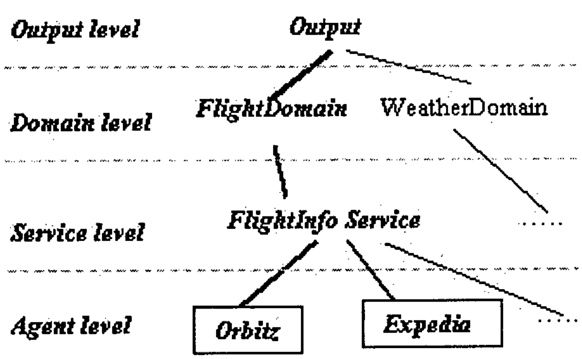

Figure 2: The hierarchical organization of the web sources

List of Personnel Associated with the Research Effort:

Craig Knoblock, PI
Kristina Lerman, Senior Research Scientist
Jose Luis Ambite, Senior Research Scientist
Maria Muslea, Research Scientist
Jean Oh, Research Scientist
Greg Barish, Graduate Research Assistant
Ion Muslea, Graduate Research Assistant
Rattapoon Tuchinda, Graduate Research Assistant

Parag Samdadiya, Graduate Research Assistant
Sheila Tejada, Graduate Research Assistant

Publications:

Jose Luis Ambite and Craig A. Knoblock.
Planning by rewriting.
Journal of Artificial Intelligence Research, 15:207--261, 2001.

Sheila Tejada, Craig A. Knoblock, and Steven Minton.
Learning object identification rules for information integration.
Information Systems, 26(8), 2001.

Hans Chalupsky, Yolanda Gil, Craig A. Knoblock, Kristina Lerman, Jean Oh, David V. Pynadath, Thomas A. Russ, and Milind Tambe.
Electric elves: Applying agent technology to support human organizations.
In *Proceedings of the Conference on Innovative Applications of Artificial Intelligence*, 2001.

Kristina Lerman, Craig A. Knoblock, and Steven Minton.
Automatic data extraction from lists and tables in web sources.
In *Proceedings of the IJCAI 2001 Workshop on Adaptive Text Extraction and Mining*, Seattle, WA, 2001.

Craig A. Knoblock, Jose Luis Ambite, Steven Minton, Cyrus Shahabi, Mohammad Kolahdouzan, Maria Muslea, Jean Oh, and Snehal Thakkar.
Integrating the world: The worldinfo assistant.
In *Proceedings of the 2001 International Conference on Artificial Intelligence (IC-AI 2001)*, Las Vegas, NV, 2001.

Craig A. Knoblock, Steven Minton, Jose Luis Ambite, Maria Muslea, Jean Oh, and Martin Frank.
Mixed-initiative, multi-source information assistants.
In *Proceedings of the World Wide Web Conference*, Hong Kong, May 2001.

Kristina Lerman, Steven Minton, and Craig A. Knoblock. Wrapper maintenance: A machine learning approach. *Journal of Artificial Intelligence Research*, 2002.

Sheila Tejada, Craig A. Knoblock, and Steven Minton. Learning domain-independent string transformation weights for high accuracy object identification. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002)*, Edmonton, Alberta, Canada, 2002.

Jose Luis Ambite, Greg Barish, Craig A. Knoblock, Maria Muslea, Jean Oh, and Steven Minton. Getting from here to there: Interactive planning and agent execution for

optimizing travel. In *Proceedings of the Fourteenth Conference on Innovative Applications of Artificial Intelligence (IAAI-2002)*, pages 862--869, Edmonton, Alberta, Canada, 2002.

Ion Muslea, Steven Minton, and Craig A. Knoblock. Active + semi-supervised learning = robust multi-view learning. In *Proceedings of the 19th International Conference on Machine Learning (ICML-2002)*, pages 435--442, Sydney, Australia, 2002.

Ion Muslea, Steven Minton, and Craig A. Knoblock. Adaptive view validation: A first step towards automatic view detection. In *Proceedings of the 19th International Conference on Machine Learning (ICML-2002)*, pages 443--450, Sydney, Australia, 2002.

Ion Muslea, Steven Minton, and Craig A. Knoblock. Adaptive view validation: A case study on wrapper induction and text classification. In *Proceedings of the AAAI-2002 Workshop on Intelligent Services Integration*, Edmonton, Alberta, Canada, 2002.

Snehal Thakkar, Craig A. Knoblock, Jose Luis Ambite, and Cyrus Shahabi. Dynamically composing web services from on-line sources. In *Proceeding of {AAAI-2002} Workshop on Intelligent Service Integration*, pages 1--7, Edmonton, Alberta, Canada, 2002.

Greg Barish and Craig A. Knoblock. Speculative execution for information gathering plans. In *Proceedings of the Sixth International Conference on Artificial Intelligence Planning and Scheduling (AIPS 2002)*, pages 259--268, Toulouse, France, 2002.

Greg Barish and Craig A. Knoblock. An efficient and expressive language for information gathering on the web. In *Proceedings of the {AIPS-2002} Workshop on Is there life after operator sequencing? -- Exploring real world planning*, pages 5--12, Tolouse, France, 2002.

Greg Barish and Craig A. Knoblock. Learning efficient value predictors for speculative plan execution. In *Proceedings of the Fifth International Workshop on the World Wide Web and Databases (WebDB 2002)*, pages 77--82, Madison, WI, 2002.

Hans Chalupsky, Yolanda Gil, Craig A. Knoblock, Kristina Lerman, Jean Oh, David V. Pynadath, ThomasA. Russ, and Milind Tambe.
Electric elves: Agent technology for supporting human organizations.
AI Magazine, 23(2):11--24, Summer 2002.

Ching-Chien Chen, Snehal Thakkar, Craig A. Knoblock, and Cyrus Shahabi.
Automatically Annotating and Integrating Spatial Datasets,
In Proceedings of the Eighth International Symposium on Spatial and Temporal Databases (SSTD 2003), Springer, Berlin, 2003.

Craig A. Knoblock, Kristina Lerman, Steven Minton, and Ion Muslea.
Accurately and reliably extracting data from the web: A machine learning approach.

In P.S. Szczepaniak, J. Segovia, J. Kacprzyk, and L.A. Zadeh, editors, *Intelligent Exploration of the Web*. Springer-Verlag, Berkeley, CA, 2003.

Craig A. Knoblock.
Deploying information agents on the web.
In Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI-2003), Acapulco, Mexico, 2003.

Ion Muslea.
Active Learning with Multiple Views.
PhD thesis, Department of Computer Science, University of Southern California, 2002.

Oren Etzioni, Craig A. Knoblock, Rattapoom Tuchinda, and Alexander Yates.
To buy or not to buy: Mining airline fare data to minimize ticket purchase price.
In Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003.

Snehal Thakkar and Craig A. Knoblock.
Efficient execution of recursive integration plans.
In Proceedings of 2003 IJCAI Workshop on Information Integration on the Web, Acapulco, Mexico, 2003.

Snehal Thakkar, Jose-Luis Ambite, and Craig A. Knoblock.
A view integration approach to dynamic composition of web services.
In Proceedings of 2003 ICAPS Workshop on Planning for Web Services, Trento, Italy, 2003.

Ion Muslea, Steven Minton, and Craig A. Knoblock.
Active learning with strong and weak views: A case study on wrapper induction.
In Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI-2003), Acapulco, Mexico, 2003.

Greg Barish and Craig A. Knoblock.
Learning value predictors for the speculative execution of information gathering plans.
In Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI-2003), Acapulco, Mexico, 2003.

Ching-Chien Chen, Craig A. Knoblock, Cyrus Shahabi, and Snehal Thakkar.
Automatically and accurately conflating satellite imagery and maps.
In Proceedings of the International Workshop on Next Generation Geospatial Information, Cambridge, MA, 2003.

Mehdi Sharifzadeh, Cyrus Shahabi, and Craig A. Knoblock.
Learning approximate thematic maps from labeled geospatial data.
In Proceedings of the International Workshop on Next Generation Geospatial Information, Cambridge, MA, 2003.

Shou de Lin and Craig A. Knoblock.
Exploiting a search engine to develop more flexible web agents.
In Proceedings of the 2003 IEEE/WIC International Conference on Web Intelligence (WI2003), pages 54--60, Halifax, Canada, 2003. IEEE Computer Society.
Best Paper Award.

Martin Michalowski, Snehal Thakkar, and Craig A. Knoblock.
Exploiting secondary sources for automatic object consolidation.
In Proceedings of the KDD'03 Workshop on Data Cleaning, Record Linkage and Object Consolidation, 2003.

Greg Barish and CraigA . Knoblock.
Combining classification and transduction for value prediction in speculative plan execution.
In Proceedings of 2003 IJCAI Workshop on Information Integration on the Web, Acapulco, Mexico, 2003.

Greg Barish.
Speculative Plan Execution for Information Agents.
PhD thesis, Department of Computer Science, University of Southern California, 2003.

Martin Michalowski, José Luis Ambite, Snehal Thakkar, Rattapoom Tuchinda, Craig A. Knoblock, and Steve Minton.
Retrieving and semantically integrating heterogeneous data from the web.
IEEE Intelligent Systems, 19(3), 2004.

Snehal Thakkar, Jose Luis Ambite, and Craig A. Knoblock.
A data integration approach to automatically composing and optimizing web services.
In Proceedings of 2004 ICAPS Workshop on Planning and Scheduling for Web and Grid Services, Whistler, BC, Canada, 2004.

Craig A. Knoblock.
Building software agents for planning, monitoring, and optimizing travel.
In Andrew J. Frew, editor, Proceedings of the Elevent International Conference on Information Technology and Travel & Tourism, Springer-Verlag, New York, 2004.

Rattapoom Tuchinda and Craig A. Knoblock.
Agent wizard: Building information agents by answering questions.
In Proceedings of Intelligent User Interfaces, Island of Madeira, Portugal, 2004.

Interactions/Transitions:

Invited Talks
Craig A. Knoblock.
Building software agents for planning, monitoring, and optimizing travel.

Elevent International Conference on Information Technology and Travel & Tourism, Cairo, Egypt, 2004.

Craig A. Knoblock.
Deploying information agents on the web.
18th International Joint Conference on Artificial Intelligence (IJCAI-2003), Acapulco, Mexico, 2003.

Craig Knoblock
Integrating Online and Geospatial Information Sources
Summer Assembly of the University Consoritum for Geographic Information Science
Monterey, CA, June 17, 2003

Craig Knoblock
Invited Tutorial on Planning and the Web
PLANET International Summer School on AI Planning
September 16-22, 2002
Halkidiki, Greece

Presentations

Hans Chalupsky, Yolanda Gil, Craig A. Knoblock, Kristina Lerman, Jean Oh, David V. Pynadath, Thomas A. Russ, and Milind Tambe.
Electric elves: Applying agent technology to support human organizations.
Conference on Innovative Applications of Artificial Intelligence, 2001. Presented by Hans Chalupsky.

Kristina Lerman, Craig A. Knoblock, and Steven Minton.
Automatic data extraction from lists and tables in web sources.
IJCAI 2001 Workshop on Adaptive Text Extraction and Mining, Seattle, WA, 2001. Presented by Kristina Lerman.

Craig A. Knoblock, Jose Luis Ambite, Steven Minton, Cyrus Shahabi, Mohammad Kolahdouzan, Maria Muslea, Jean Oh, and Snehal Thakkar.
Integrating the world: The worldinfo assistant.
2001 International Conference on Artificial Intelligence (IC-AI 2001), Las Vegas, NV, 2001. Presented by Jose Luis Ambite.

Craig A. Knoblock, Steven Minton, Jose Luis Ambite, Maria Muslea, Jean Oh, and Martin Frank.
Mixed-initiative, multi-source information assistants.
World Wide Web Conference, Hong Kong, May 2001. Presented by Craig Knoblock
Craig Knoblock. Planning for Information Integration on the Web. Invited talk at the AIPS-2002 Workshop on Is There Live Beyond Operator Sequencing? -- Exploring Real World Planning.

Sheila Tejada, Craig A. Knoblock, and Steven Minton. Learning domain-independent string transformation weights for high accuracy object identification. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002)*, Edmonton, Alberta, Canada, 2002.
Presented by Sheila Tejada.

Jose Luis Ambite, Greg Barish, Craig A. Knoblock, Maria Muslea, Jean Oh, and Steven Minton. Getting from here to there: Interactive planning and agent execution for optimizing travel. In *Proceedings of the Fourteenth Conference on Innovative Applications of Artificial Intelligence (IAAI-2002)*, pages 862--869, Edmonton, Alberta, Canada, 2002.
Presented by Jose Luis Ambite.

Ion Muslea, Steven Minton, and Craig A. Knoblock. Active + semi-supervised learning = robust multi-view learning. In *Proceedings of the 19th International Conference on Machine Learning (ICML-2002)*, pages 435--442, Sydney, Australia, 2002.
Presented by Ion Muslea.

Ion Muslea, Steven Minton, and Craig A. Knoblock. Adaptive view validation: A first step towards automatic view detection. In *Proceedings of the 19th International Conference on Machine Learning (ICML-2002)*, pages 443--450, Sydney, Australia, 2002.
Presented by Ion Muslea.

Ion Muslea, Steven Minton, and Craig A. Knoblock. Adaptive view validation: A case study on wrapper induction and text classification. In *Proceedings of the AAAI-2002 Workshop on Intelligent Services Integration*, Edmonton, Alberta, Canada, 2002.
Presented by Ion Musela.

Snehal Thakkar, Craig A. Knoblock, Jose Luis Ambite, and Cyrus Shahabi. Dynamically composing web services from on-line sources. In *Proceeding of {AAAI-2002} Workshop on Intelligent Service Integration*, pages 1--7, Edmonton, Alberta, Canada, 2002.
Presented by Snehal Thakkar.

Greg Barish and Craig A. Knoblock. Speculative execution for information gathering plans. In *Proceedings of the Sixth International Conference on Artificial Intelligence Planning and Scheduling (AIPS 2002)*, pages 259--268, Toulouse, France, 2002.
Presented by Greg Barish.

Greg Barish and Craig A. Knoblock. An efficient and expressive language for information gathering on the web. In *Proceedings of the {AIPS-2002} Workshop on Is there life after operator sequencing? -- Exploring real world planning*, pages 5--12, Tolouse, France, 2002.
Presented by Greg Barish.

Greg Barish and Craig A. Knoblock. Learning efficient value predictors for speculative plan execution. In *Proceedings of the Fifth International Workshop on the World Wide Web and Databases (WebDB 2002)*, pages 77--82, Madison, WI, 2002.
Presented by Greg Barish.

Ching-Chien Chen, Snehal Thakkar, Craig A. Knoblock, and Cyrus Shahabi.
Automatically Annotating and Integrating Spatial Datasets,
Eighth International Symposium on Spatial and Temporal Databases (SSTD 2003), Springer, Berlin, 2003.
Presented by Cyrus Shahabi

Oren Etzioni, Craig A. Knoblock, Rattapoom Tuchinda, and Alexander Yates.
To buy or not to buy: Mining airline fare data to minimize ticket purchase price.
Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003.
Presented by Oren Etzioni

Snehal Thakkar and Craig A. Knoblock.
Efficient execution of recursive integration plans.
2003 IJCAI Workshop on Information Integration on the Web, Acapulco, Mexico, 2003.
Presented by Snehal Thakkar.

Snehal Thakkar, Jose-Luis Ambite, and Craig A. Knoblock.
A view integration approach to dynamic composition of web services.
2003 ICAPS Workshop on Planning for Web Services, Trento, Italy, 2003.
Presented by Snehal Thakkar.

Ion Muslea, Steven Minton, and Craig A. Knoblock.
Active learning with strong and weak views: A case study on wrapper induction.
In Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI-2003), Acapulco, Mexico, 2003.
Presented by Steve Minton.

Greg Barish and Craig A. Knoblock.
Learning value predictors for the speculative execution of information gathering plans.
18th International Joint Conference on Artificial Intelligence (IJCAI-2003), Acapulco, Mexico, 2003.
Presented by Greg Barish.

Ching-Chien Chen, Craig A. Knoblock, Cyrus Shahabi, and Snehal Thakkar.
Automatically and accurately conflating satellite imagery and maps.
In Proceedings of the International Workshop on Next Generation Geospatial Information, Cambridge, MA, 2003.
Preseneted by Ching-Chien Chen.

Mehdi Sharifzadeh, Cyrus Shahabi, and Craig A. Knoblock.
Learning approximate thematic maps from labeled geospatial data.
In Proceedings of the International Workshop on Next Generation Geospatial Information, Cambridge, MA, 2003.
Presented by Mehdi Sharifzadeh.

Shou de Lin and Craig A. Knoblock.
Exploiting a search engine to develop more flexible web agents.
In Proceedings of the 2003 IEEE/WIC International Conference on Web Intelligence (WI2003), pages 54--60, Halifax, Canada, 2003. IEEE Computer Society.
Presented by Shou-de Lin.

Martin Michalowski, Snehal Thakkar, and Craig A. Knoblock.
Exploiting secondary sources for automatic object consolidation.
In Proceedings of the KDD'03 Workshop on Data Cleaning, Record Linkage and Object Consolidation, 2003.
Presented by Martin Michalowski.

Greg Barish and CraigA . Knoblock.
Combining classification and transduction for value prediction in speculative plan execution.
In Proceedings of 2003 IJCAI Workshop on Information Integration on the Web, Acapulco, Mexico, 2003.
Presented by Greg Barish.

Snehal Thakkar, Jose Luis Ambite, and Craig A. Knoblock.
A data integration approach to automatically composing and optimizing web services.
In Proceedings of 2004 ICAPS Workshop on Planning and Scheduling for Web and Grid Services, Whistler, BC, Canada, 2004.
Presented by Snehal Thakkar.

Consultative and Advisory Functions

- Craig Knoblock participated in the AFOSR annual meeting in Ithaca, NY, May, 2001.
- Craig Knoblock visited the National Imagery and Mapping Agency and presented his research on integrating open source data with geospatial data, June, 2001.
- Craig Knoblock attended the DARPA CoABS PI Meeting in Nashua, NH, July, 2001.
- Craig Knoblock attended the DARPA Active Templates PI Meeting in Washington,DC, November, 2002.
- Craig Knoblock attended the DARPA EELD PI Meetings in Washington, DC in October, 2001 and San Francisco, CA, June, 2002.
- Craig Knoblock attended the DARPA EELD PI Meetings in Washington, DC, November, 2001 and in Savannah, GA, May, 2002.
- Craig Knoblock attended the DARPA CoABS PI Meeting in Washington,DC, January, 2002.
- Craig Knoblock participated in the AFOSR annual meeting in Syracuse, NY, June, 2002.

- Craig Knoblock presented research results at a meetings with Doug Dyer, Warren Knouff, Fred Bobbitt, Terry Sullivan, and other personnel at Ft. Bragg in Fayetteville, NC on July 10, 2002.
- Craig Knoblock participated in the AFOSR annual meeting in Syracuse, NY, June, 2003.
- Craig Knoblock presented his work on Geospatial Data Integration at AFRL in Rome, NY in May, 2004. Visit was hosted by John Salerno.

Transitions

- The Heracles system, which builds on work funded by AFOSR, has been deployed within Special Operations.
- Fetch Technologies, Inc (www.fetch.com) continues to use the Theseus Agent Execution System, which was developed under our previous AFOSR grant. Fetch was recently awarded an Air Force SBIR Phase II Grant to apply the Theseus technology to the Joint Battlespace Infosphere.
- Fetch Technologies and USC were jointly awarded a AFOSR STTR Phase II Grant to transition our previous work on object identification into a commercial product.

Discoveries/Inventions/Patent Disclosures

- Filed Utility Patent Application, INFORMATION AGENT SYSTEM("THESEUS") Greg Barish, Craig Knoblock, Steve Minton, and John Daniel Rosenberry, Serial #:09/707,147, Date filed:11/3/00, U.S.
- Filed Utility Patent Application, CO-TESTING, Ion Muslea, Craig Knoblock, and Steve Minton, Serial#:Pending, Date filed:4/6/01, U.S.
- Filed Invention Disclosure, No: 3411, Mining Airline Fare Data to Minimize Ticket Purchase Price, Patent Application in process.
- Filed Invention Disclosure, No: 3330, A Technique for The Speculative Execution of Streaming Dataflow Plans
- Filed Invention Disclosure, No: 3236, Constraint-based Information Gathering and Integration

Honors/Awards

- Craig Knoblock was promoted to Senior Project Leader, May, 2001
- Craig Knoblock was also elected to the AAAI Executive Council, July 2001.
- Craig Knoblock was elected to Treasurer of the ICAPS Council (International Conference on Automated Planning and Scheduling), 2002
- Craig Knoblock was elected President-elect of the ICAPS Council, 2004
- Craig Knoblock gave an invited talk at the International Joint Conference on Artificial Intelligence on the research supported by this grant.
- Craig Knoblock was named a Fellow of the American Association of Artificial Intelligence